EDF TESTS FOR NORMALITY IN LINEAR MODELS AFTER A BOX-COX TRANSFORMATION

by

Gemai Chen

Richard Lockhart

Michael A. Stephens

TECHNICAL REPORT No. 472

JULY 2, 1993

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

EDF Tests for Normality in Linear Models after a Box-Cox Transformation

Gemai Chen, Richard Lockhart and Michael A. Stephens

Simon Fraser University, B.C. Canada

Summary

The Box-Cox transformation procedure has been used extensively in data analysis, for example in regression, where the response variable is subjected to a suitable power transformation so that the standard normal-theory linear regression models can be fitted to the transformed values. In this paper, distribution theory is developed for a family of EDF statistics, including the Anderson-Darling statistic A^2 and the Cramér-von Mises statistic W^2, so that these statistics can be used to test for normality in the linear model after applying the Box-Cox transformation. A table of asymptotic critical points is given for A^2 and W^2, and numerical examples are given to illustrate the use of the table.

Keywords: LINEAR REGRESSION, NON-LINEAR REGRESSION, MAXIMUM LIKELIHOOD ESTIMATION, TRANSFORMATIONS TO NORMALITY

1 Introduction

The Box-Cox transformation procedure has been used extensively in regression analysis, in which the response variable is subjected to a suitable power transformation so that the standard normal-theory linear regression model can be fitted to the transformed responses.

Let $Y_1, \ldots, Y_n$ be positive independent random variables denoting responses to variables X. For a real number λ, the Box-Cox power transformation family is

$$Y_i(\lambda) = \begin{cases} (Y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log Y_i & \text{if } \lambda = 0. \end{cases} \tag{1.1}$$

The Box-Cox transformation is used to find a suitable λ so that, after transformation, the following linear model is approximately applicable,

$$Y(\lambda) \approx X\beta + \sigma\varepsilon, \tag{1.2}$$

where $X = (x_{ij})$ is a known $n \times p$ matrix of constants, $\beta = (\beta_1, \ldots, \beta_p)^t$ are unknown regression parameters (a column vector), σ is an unknown positive constant, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ are independent and identically distributed standard normal random variables, and $Y(\lambda) = (Y_1(\lambda), \ldots, Y_n(\lambda))^t$; superscript t denotes transpose.

An important part of this model is the assumption that the ϵ_i are i. i. d. $N(0,1)$, and in this paper we propose tests for this assumption. The tests will be based on the empirical distribution function (EDF) of the estimated residuals, and we give tables for the Cramér-von Mises statistic W^2 and the Anderson-Darling statistic A^2.

We begin by making two important comments. Firstly, a test for normality after regression, without the Box-Cox transformation, and using the estimated residuals, is known to be the same asymptotically as a one-sample test for which the mean and variance must be estimated; see Stephens (1986, Section 4.8.5) for the procedure. Here the situation is different, essentially because λ must also be estimated, and one might expect that the tables of percentage points for the test will depend, even asymptotically, on the true values of the various parameters. However, we show that the problem can be reduced so that, in most practical circumstances, the tables depend on only one new parameter g, to be given in Section 2; g is a function of the estimated regression parameters and of $\hat{\sigma}$, the estimate of σ.

Secondly, we observe that in model (1.2), the parameters are usually estimated as though there were no restraint on the $Y(\lambda)$. However, (1.2) cannot be precisely correct except when $\lambda = 0$, since the right hand side can take on any value, but the left hand side must be greater than $-1/\lambda$. In practice, this restriction will make very little difference to the algebra of fitting model (1.2); we discuss this issue in Section 2.

Nevertheless, for the purpose of developing a sound theory for the tests of fit, it is necessary to use the correct density and distribution of $Y_i, i = 1, 2, \ldots, n$. The density is the following truncated normal distribution (see, for example, Poirier, 1978)

$$f(y_i; \lambda, \mu_i, \sigma^2) = \begin{cases} \sigma^{-1}\phi\left[\{(y_i^{\lambda} - 1)/\lambda - \mu_i\}/\sigma\right] y_i^{\lambda-1}/\Phi(\delta_i), & \text{if } \lambda > 0, \\ \sigma^{-1}\phi\{(\log y_i - \mu_i)/\sigma\} y_i^{-1}, & \text{if } \lambda = 0, \\ \sigma^{-1}\phi\left[\{(y_i^{\lambda} - 1)/\lambda - \mu_i\}/\sigma\right] y_i^{\lambda-1}/\Phi(-\delta_i), & \text{if } \lambda < 0. \end{cases} \quad (1.3)$$

In (1.3), $y_i > 0$; in addition, $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of a standard normal random variable, $\mu_i = x_i^t\beta$, with x_i^t denoting the i^{th} row of X, and $\delta_i = (\mu_i + 1/\lambda)/\sigma$. Note that in model (1.3) $E(Y_i) \neq \mu_i$, so that model (1.3) is a non-linear model in β.

The plan of this paper is the following. Section 2 discusses parameter estimation in model (1.3) and presents a family of EDF statistics for testing the fit of the model. In particular, we consider the effect of estimating λ, and introduce the parameter g which is the argument of the tables used with the tests. Section 3 illustrates the use of the EDF tests with real data. Section 4 contains the theory for the EDF tests introduced in section 2, and in Section 5 the accuracy of the tests is examined. Technical details are given in an Appendix.

2 Parameter Estimation and EDF Tests of Fit

2.1 Parameter Estimation

We first introduce ν for σ^2 and then use maximum likelihood to estimate parameters λ, β and ν. Denote by L the log-likelihood function of a random sample $Y_1, \ldots, Y_n$ based on model (1.3); then except for a constant,

$$L = \begin{cases} -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^n\{(y_i^\lambda - 1)/\lambda - \mu_i\}^2 & \\ +(\lambda-1)\sum_{i=1}^n \log y_i - \sum_{i=1}^n \log\Phi(\delta_i), & \text{if } \lambda > 0, \\ -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^n\{\log y_i - \mu_i\}^2 & \\ -\sum_{i=1}^n \log y_i, & \text{if } \lambda = 0, \\ -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^n\{(y_i^\lambda - 1)/\lambda - \mu_i\}^2 & \\ +(\lambda-1)\sum_{i=1}^n \log y_i - \sum_{i=1}^n \log\Phi(-\delta_i), & \text{if } \lambda < 0. \end{cases} \tag{2.1}$$

Because $Y_i(\lambda)$ defined by (1.1) is differentiable with respect to λ, L is differentiable with respect to λ, β and ν. Thus, for $\lambda > 0$ the likelihood equations are

$$\frac{\partial L}{\partial \beta_k} = \nu^{-1}\sum_{i=1}^n[(y_i^\lambda - 1)/\lambda - \mu_i]x_{ik} - \sum_{i=1}^n[\phi(\delta_i)/\Phi(\delta_i)][x_{ik}/\sqrt{\nu}] = 0, \; k = 1, \ldots, p \tag{2.2}$$

$$\frac{\partial L}{\partial \nu} = -n/(2\nu) + (2\nu^2)^{-1}\sum_{i=1}^n[(y_i^\lambda - 1)/\lambda - \mu_i]^2 + \sum_{i=1}^n[\phi(\delta_i)/\Phi(\delta_i)][\delta_i/(2\nu)] = 0, \tag{2.3}$$

$$\begin{aligned} \frac{\partial L}{\partial \lambda} &= -(\lambda^2\nu)^{-1}\sum_{i=1}^n[(y_i^\lambda - 1)/\lambda - \mu_i](\lambda y_i^\lambda \log y_i - y^\lambda + 1) \\ &\quad + \sum_{i=1}^n[\phi(\delta_i)/\Phi(\delta_i)][1/(\lambda\nu)] + \sum_{i=1}^n \log y_i = 0. \end{aligned} \tag{2.4}$$

Similar likelihood equations can be found for $\lambda < 0$ and $\lambda = 0$.

It does not seem possible to find closed-form maximum likelihood estimators for β, ν and λ from the likelihood function directly, or from the likelihood equations. Therefore, iterative numerical methods are necessary.

The usual Box-Cox approach to parameter estimation gives a log-likelihood function which is related to L above. We denote by l_{BC} the log-likelihood function of model (1.2) discussed by Box and Cox (1964). Then, except for a constant,

$$l_{BC} = -(n/2)\log\nu - (2\nu)^{-1}\sum_{i=1}^{n}\{(y_i^{\lambda} - 1)/\lambda - \mu_i\}^2 + (\lambda - 1)\sum_{i=1}^{n}\log y_i. \tag{2.5}$$

The Box-Cox method maximizes l_{BC} over β and ν with λ kept fixed; the result is a function of λ alone that is given by

$$l_{BC}(\lambda) = -(n/2)\log\tilde{\nu}_{BC}(\lambda) - (n/2) + (\lambda - 1)\sum_{i=1}^{n}\log y_i, \tag{2.6}$$

where $n\tilde{\nu}_{BC}(\lambda) = Y(\lambda)^t(I - X(X^tX)^{-1}X^t)Y(\lambda)$ is the residual sum of squares from regressing $Y(\lambda)$ on X. The final value $\tilde{\lambda}$ of λ for subsequent analysis is determined by maximizing $l_{BC}(\lambda)$ over λ, and the estimates for β and ν are then given by

$$\begin{aligned} \tilde{\beta} &= (X^tX)^{-1}X^tY(\tilde{\lambda}), & (2.7) \\ \tilde{\nu} &= \frac{1}{n}Y(\tilde{\lambda})(I - X(X^tX)^{-1}X^t)Y(\tilde{\lambda}). & (2.8) \end{aligned}$$

Note that the residual sum of squares is usually divided by $n - p$ to obtain an estimate of ν, where p is the number of regression parameters. The examples given in the next section will follow this convention.

It is useful to compare the results obtained by using l_{BC} rather than L to estimate the parameters. It can be seen by comparing (2.5) to (2.1) that $l_{BC} \approx L$ if $-\sum_{i=1}^{n}\log\Phi(\delta_i) \approx 0$, if $\lambda > 0$, or $-\sum_{i=1}^{n}\log\Phi(-\delta_i) \approx 0$, if $\lambda < 0$. This happens if (1) λ is close to zero, or (2) μ_i's (or $-\mu_i$'s) are large, or (3) ν is small. In practice, one or more of these conditions often holds. Consider, for example, the case $\lambda > 0$. Suppose $\delta_n^+ = \min_{1\leq i\leq n}\{\delta_i\}$; if $\delta_n^+ > \Phi^{-1}(e^{-c/n})$ for a positive constant c, then

$$-\sum_{i=1}^{n}\log\Phi(\delta_i) \leq -n\log\Phi(\delta_n^+) < c.$$

For example, suppose $c = 0.01$ and $n = 50$; then $\Phi^{-1}(e^{-0.01/50}) = 3.54$. Thus, if the minimum $\delta_i = \min(\mu_i + 1/\lambda)/\sigma$ has a value 3.54, then truncating the left tail as in (1.3) cuts off less than 0.01 from the log-likelihood. A similar result holds for the case when $\lambda < 0$. It often happens that $-\sum_{i=1}^{n}\log\phi(\delta_i)$ (or, for $\lambda < 0$, $-\sum_{i=1}^{n}\log\phi(-\delta_i)$) is small, so that $l_{BC} \approx L$, and then use of l_{BC} to estimate parameters from a given set of data will yield almost the same results as use of the likelihood L.

2.2 EDF Tests of Fit

To test for goodness-of-fit when fitting model (1.3) to data, we use the well-known Cramér-von Mises family of EDF statistics. For the present case, let $\gamma = (\lambda, \beta^t, \nu)^t$ and let $\hat{\gamma} = (\hat{\lambda}, \hat{\beta}^t, \hat{\nu})^t$ be the maximum likelihood estimate for γ. The cumulative distribution function for Y_i in model (1.3) is given by

$$F_i(y_i; \gamma) = \begin{cases} \{\Phi(y_i^*) + \Phi(\delta_i) - 1\}/\Phi(\delta_i), & \text{if } \lambda > 0, \\ \Phi\{(\log y_i - \mu_i)/\sqrt{\nu}\}, & \text{if } \lambda = 0, \\ \Phi(y_i^*)/\Phi(-\delta_i), & \text{if } \lambda < 0, \end{cases} \tag{2.9}$$

where $\mu_i = x_i^t\beta$, $\delta_i = (\mu_i + 1/\lambda)/\sqrt{\nu}$, $y_i^* = ((y_i^\lambda - 1)/\lambda - \mu_i)/\sqrt{\nu}$. Now for each i, let $v_i = F_i(y_i; \hat{\gamma})$ and let the empirical distribution function of the v_i's be

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1[v_i \leq t], \ (0 \leq t \leq 1), \tag{2.10}$$

where $1[A] = 1$ if A is true, otherwise, $1[A] = 0$. The EDF statistics are based on the discrepancies between $\hat{F}_n(t)$ and $F(t) \equiv t$ $(0 \leq t \leq 1)$, namely,

$$Q_n = n\int_0^1 \{\hat{F}_n(t) - t\}^2 \psi(t) dt, \tag{2.11}$$

where $\psi(t) > 0$ is a suitable known weight function. As special cases, the Cramér-von Mises statistic W^2 is obtained when $\psi(t) \equiv 1$, and the Anderson-Darling statistic A^2 is obtained when $\psi(t) = \{t(1-t)\}^{-1}$. Let $v_{(1)} < v_{(2)} < \cdots < v_{(n)}$ be the order statistics of the v_i. Statistics W^2 and A^2 are then given by

$$W^2 = \sum_{i=1}^{n}\left\{v_{(i)} - \frac{2i-1}{2n}\right\}^2 + \frac{1}{12n}, \tag{2.12}$$

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\{(2i-1)\log v_{(i)} + (2n+1-i)\log(1 - v_{(i)})\}. \tag{2.13}$$

2.3 The Goodness-of-Fit Test Procedure

Suppose that the model matrix X is of the form $X = (1_n \ V)$ where 1_n is an n by 1 vector of 1's. Center matrix V into U by subtracting from each column the mean of that column. Let u_i^t denote the i^{th} row of U. Denote the regression parameters β by $\beta = (\beta_1, \ldots, \beta_p)^t = (\beta_1, \theta^t)^t$, where $\theta^t = (\beta_2, \beta_3, \ldots, \beta_p)$; then when no covariates are involved, $\beta = \beta_1 = \mu$, say, where μ is the grand mean.

To perform a goodness-of-fit test of H_0: model (1.3) fits the data, the following steps are taken:

(a) Find $\hat{\lambda}$, $\hat{\beta}$ and $\hat{\nu}$ as described above,

(b) Compute $v_i = F_i(y_i; \hat{\gamma})$ according to (2.9) with the true parameters replaced by their estimates. In practice, the use of $v_i = \Phi(\hat{y}_i^*)$, where $\hat{y}_i^* = (\{(y_i^{\hat{\lambda}} - 1)/\lambda\} - \hat{\mu}_i)/\hat{\sigma}$, $\hat{\sigma} = \sqrt{\hat{\nu}}$ will almost always give the same test result.

(c) Put the v_i in ascending order, and calculate W^2 or A^2 according to (2.12) or (2.13), respectively,

(d) Find $\eta_i = u_i^t\hat{\theta}/\sqrt{\hat{\nu}}$, and obtain a quantity g defined by

$$g = 6 + \frac{1}{n}\sum_{i=1}^{n}\left\{8\eta_i^2 + \eta_i^4\right\} - \left(\frac{1}{n}\sum_{i=1}^{n}\eta_i^2\right)^2 - a_n(\frac{1}{n}U^tU)^{-1}a_n^t, \tag{2.14}$$

where $a_n = n^{-1}\sum_{i=1}^{n}\eta_i^2 u_i^t$.

(e) Enter Table 1 with the value of $1/g$, and reject H_0 at significance level α if the test statistic exceeds the corresponding upper α-percentile given in Table 1.

The entries in Table 1 are the upper percentiles, for the appropriate $1/g$, of the asymptotic distributions of W^2 and A^2, respectively, as $n \to \infty$, and as $|(\beta_1 + 1/\lambda)/\sigma| \to \infty$. Note that the upper percentiles corresponding to $1/g = 0$ are the upper percentiles for testing goodness-of-fit of linear models *without* taking any Box-Cox transformation (see Stephens (1986), Section 4.8.5). The percentiles for $1/g$ differ increasingly from these values as $1/g$ grows larger. When $g = 6$ the situation corresponds to the case where there is no regression, and the $y_i(\lambda)$ are simply a transformed random sample.

In principle, the tabulated distributions could be expected to depend on the various values of the parameters θ and λ, but in fact the effect of estimating these parameters is all contained

Table 1: Upper percentiles of the asymptotic distributions of W^2 and A^2 for testing Box-Cox transformations when $|(\beta_1 + 1/\lambda)/\sigma| \to \infty$ and $n \to \infty$.

		Upper Percentiles						
		α						
Statistics	$1/g$	0.50	0.25	0.20	0.15	0.10	0.05	0.01
W^2	0	0.0508	0.0739	0.0812	0.0915	0.1036	0.1260	0.1787
	1/1000	0.0508	0.0738	0.0810	0.0905	0.1031	0.1258	0.1785
	1/600	0.0508	0.0738	0.0810	0.0903	0.1031	0.1257	0.1783
	1/400	0.0507	0.0737	0.0809	0.0902	0.1030	0.1256	0.1781
	1/200	0.0506	0.0736	0.0807	0.0899	0.1028	0.1251	0.1773
	1/100	0.0504	0.0731	0.0802	0.0894	0.1022	0.1243	0.1759
	1/60	0.0501	0.0726	0.0796	0.0887	0.1011	0.1231	0.1740
	1/40	0.0498	0.0719	0.0787	0.0878	0.1002	0.1217	0.1716
	1/20	0.0487	0.0700	0.0766	0.0851	0.0970	0.1175	0.1649
	1/10	0.0463	0.0660	0.0721	0.0800	0.0909	0.1097	0.1530
	1/6	0.0428	0.0608	0.0663	0.0736	0.0836	0.1007	0.1406
A^2	0	0.3405	0.4702	0.5100	0.5607	0.6318	0.7530	1.0375
	1/1000	0.3403	0.4697	0.5094	0.5601	0.6310	0.7520	1.0351
	1/600	0.3400	0.4693	0.5090	0.5596	0.6304	0.7512	1.0339
	1/400	0.3398	0.4689	0.5085	0.5590	0.6297	0.7504	1.0326
	1/200	0.3392	0.4677	0.0571	0.5574	0.6277	0.7476	1.0281
	1/100	0.3378	0.4653	0.5043	0.5541	0.6236	0.7422	1.0187
	1/60	0.3359	0.4620	0.5005	0.5496	0.6182	0.7351	1.0007
	1/40	0.3335	0.4578	0.4958	0.5441	0.6115	0.7262	0.9928
	1/20	0.3262	0.4454	0.4817	0.5277	0.5918	0.7004	0.9518
	1/10	0.3106	0.4202	0.4537	0.4958	0.5546	0.6537	0.8820
	1/6	0.2871	0.3880	0.4186	0.4575	0.5117	0.6035	0.8168

in the one estimated parameter g. The accuracy of the above tests is discussed in Sections 4 and 5.

3 Examples

Three examples are given below to illustrate the use of Table 1. These examples deal with three typical situations where λ is positive, close to zero, and negative.

Example 1. Textile Data. Table 4 of Box and Cox (1964) contains the result of a single replicate of a 3^3 factorial experiment. The response y is the cycles to failures of worsted yarn, and the three explanatory variables assume three different levels each; see Box and Cox (1964) for details.

Three main effect linear models are fitted to the data. The first model uses y directly. The second model transforms y according to (1.1) and the third model uses the log transformation since the estimate of λ is very close to 0. Parameter estimates are obtained by directly maximizing the log-likelihood function and by applying the Box-Cox transformation procedure; the results are practically the same, and $\hat{\sigma}$ and $\hat{\lambda}$ are given in Table 2. The much smaller values of W^2 and A^2 show that the transformed models are much better fits to the data.

Table 2: EDF tests of fit for three main effect linear models, textile data, Example 1.

Model	Parameter Estimates			Modified EDF (P-value)	
	$\hat{\sigma}$	$\hat{\lambda}$	g	A^2	W^2
y:	488.2	—	—	1.3523 (<0.01)	0.2364 (<0.01)
$y(\lambda)$:	0.126	−0.059	1042	0.3372 (>0.50)	0.0495 (>0.50)
$\log y$:	0.186	0	981	0.2480 (>0.50)	0.0323 (>0.50)
$y(\lambda)$:	Minimum of $-\delta_i = 82.577$			$-\sum_{i=1}^{27} \log \Phi(-\delta_i) \approx 0$	

Example 2. Tree Data The tree data in the *Minitab Student Handbook* (Ryan, Joiner and Ryan, 1976, page 278) are analyzed here. The heights (x_1), the diameters (x_2) at 4.5 ft

above ground level and the volumes (y) were measured for a sample of 31 black cherry trees in the Allegheny National Forest, Pennsylvania. The data were collected to determine an easy way of estimating the volume of a tree based on its height and diameter.

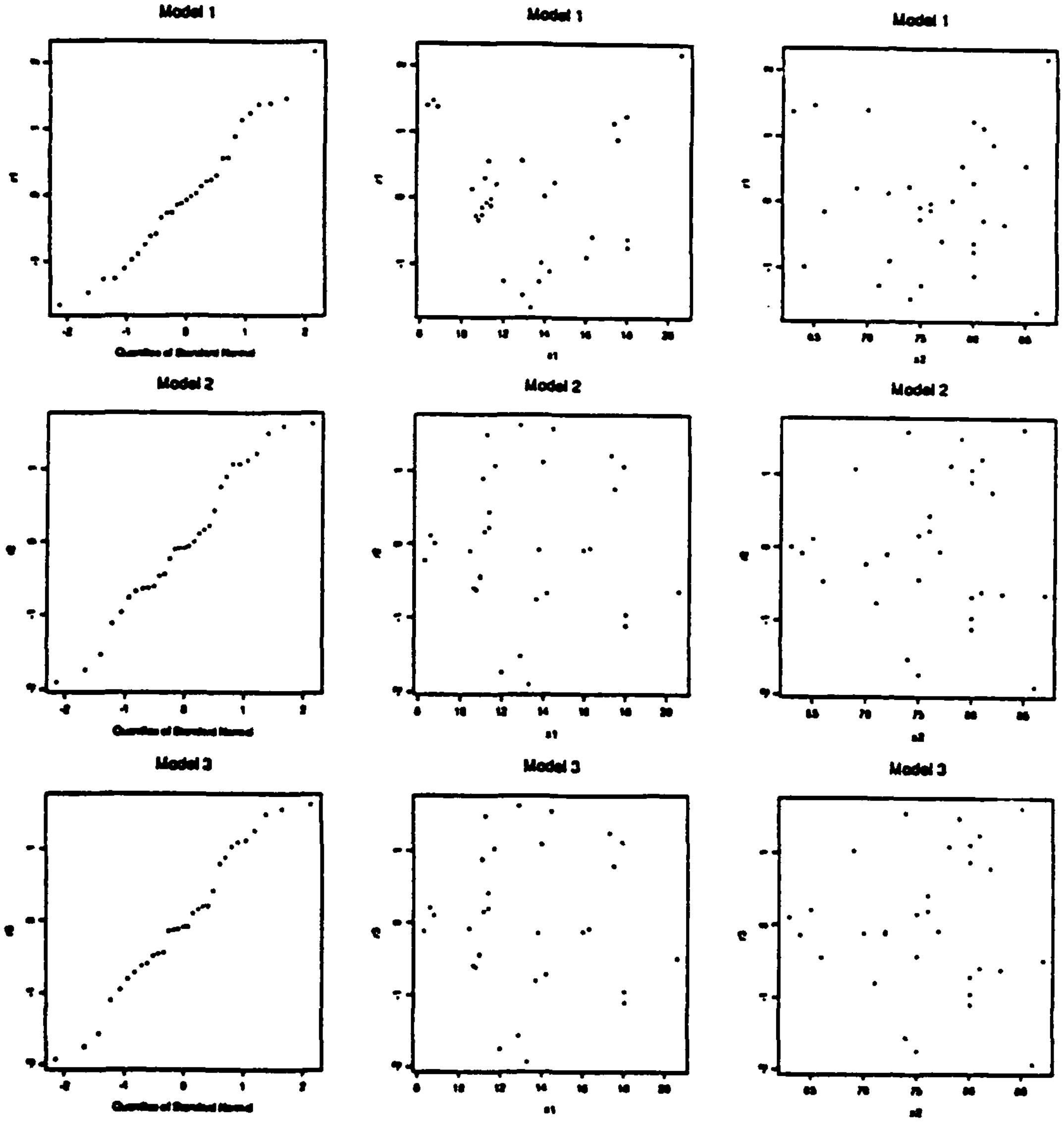
Again, three linear models are fitted to the data, using y, $y(\lambda)$ and $y(\frac{1}{3})$, where $\frac{1}{3}$ is chosen from the dimension of volume versus length. Parameter estimates were obtained by directly maximizing the log-likelihood function and by applying the Box-Cox transformation procedure; the estimates are again virtually the same. Table 3 contains the results.

Table 3: EDF tests of fit for three straight line models, tree data, Example 2.

Model	Parameter Estimates			Modified EDF (P-value)	
	$\hat{\sigma}$	$\hat{\lambda}$	g	A^2	W^2
1 y:	3.882	—	—	0.2482 (>0.50)	0.0361 (> 0.50)
2 $y(\lambda)$:	0.227	0.307	2877	0.2925 (>0.50)	0.0450 (>0.50)
3 $y(\frac{1}{3})$:	0.249	$\frac{1}{3}$	2894	0.2735 (>0.50)	0.0407 (>0.50)
$y(\lambda)$:	Minimum of $\delta_i = 29.22$			$-\sum_{i=1}^{25} \log \Phi(\delta_i) \approx 0$	

In this example, the Box-Cox estimate $\hat{\lambda} = 0.307$ is close to the estimate 1/3 derived from dimensional considerations. All of the three models pass the EDF tests easily, with the untransformed data giving slightly better values of W^2 and A^2. However, a close look at residual plots (Figure 1) suggests that the transformed models are better than the untransformed one. It appears that normality is sacrificed a little in order to obtain overall better fits.

Figure 1: Q-Q plots and residual plots for example 2



Q-Q plots and plots of residuals against regressors x_1 and x_2 for example 2. Model 1 uses y, model 2 uses $y(\lambda)$, and model 3 uses $y(\frac{1}{3})$; r_i denotes the standardized residuals.

Example 3. Biological Data. In Table 1 of Box and Cox (1964), the entries are the survival times (unit is 10 hours) of animals in a 3 × 4 completely randomized factorial experiment. The factors are Poison Content with three levels and Treatment with four levels.

Three main effect models are fitted to the data as in the two previous examples; the third model (with $\lambda = -1$, the closest integer to $\hat{\lambda}$) is included for comparison. Table 4 clearly shows that the power transformation improves the model fit considerably.

Table 4: EDF tests of fit for three main effect linear models, biological data, Example 3.

Model	Parameter Estimates			Modified EDF (P-value)	
	$\hat{\sigma}$	$\hat{\lambda}$	g	A^2	W^2
y:	0.1582	—	—	1.0373 (<0.05)	0.1572 (<0.05)
$y(\lambda)$:	0.3916	−0.75	62	0.1974 (>0.50)	0.0281 (>0.50)
$y(-1)$:	0.4931	−1.00	64	0.2861 (>0.50)	0.0387 (>0.50)
$y(\lambda)$:	Minimum of $-\delta_i = 3.667$			$-\sum_{i=1}^{48} \log \Phi(-\delta_i) = 0.0005$	

4 Theory of the Tests

4.1 The case $\lambda = 0$

In order to obtain and use the asymptotic distributions of A^2 and W^2, a key step is to show that the (estimated) empirical process

$$\hat{Y}_n(t) = \sqrt{n}(\hat{F}_n(t) - t) \tag{4.1}$$

converges weakly to a Gaussian process with zero mean and a manageable covariance function. In Theorem 4.1, we cover the situation when $\lambda = 0$. The first part of the theorem gives the asymptotic distribution of $\hat{\theta}$, the m. l. e. of θ.

Theorem 4.1 *In model (1.3), suppose that*

(A) $X = (1_n\ U)$ *is such that* $1_n^t U = 0$*, where* 1_n *is an* $n \times 1$ *vector of 1's,*

(B) $E = \lim_{n\to\infty} n^{-1} X^t \mu^2$ *and* $b = \lim_{n\to\infty} n^{-1} 1_n^t \mu^4$ *exist for any* $\beta \in \Omega$*, where* Ω *is an open convex subset of* R^p*,* $\mu = X\beta$*,* μ^k *is an* $n \times 1$ *vector with its* i^{th} *component equal to* $(x_i^t\beta)^k$*,* $k = 2, 4$*,*

(C) $\Delta = \lim_{n\to\infty} n^{-1} X^t X$ *exists and is positive definite,*

(D) *there are constants* M_1 *and* M_2 *such that for any* n *and any* $i = 1, \ldots, n$*,*

$$\frac{1}{n}\sum_{i=1}^{n} \max_{1\le j\le p} |x_{ij}| \le M_1,$$

$$\frac{1}{\sqrt{n}} \max_{1\le j\le p} |x_{ij}| \le M_2,$$

(E) $c_1 = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \eta_i^2$*,* $c_2 = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \eta_i^4$*, and* $a = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \eta_i^2 u_i^t$ *exist, where* $\eta_i = u_i^t(\beta_2, \ldots, \beta_p)^t/\sigma$ *and* u_i^t *is the* i^{th} *row of* U*,*

then

(1) *when $\lambda = 0$, the maximum likelihood estimate $\hat{\theta}$ of $\theta \equiv (\beta_2, \beta_3, \ldots, \beta_p)^t$ is asymptotically normal, that is,*

$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \Gamma)$, *where*

$$\Gamma = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix}^{-1},$$

with

$$A = (4\nu)^{-1}(7\nu^2 + 10\nu\beta^t\Delta\beta + b),$$

$$B = (-D/2 - E^t/(2\nu),\ -\beta_1/\nu),$$

$$C = \begin{pmatrix} \Delta/\nu & 0 \\ 0^t & 1/(2\nu^2) \end{pmatrix},$$

where $D^t = (1, 0, \ldots, 0)^t$ is a $p \times 1$ vector with its first component equal to 1 and all the other components equal to 0;

(2) *when $\lambda = 0$, the (estimated) empirical process $\hat{Y}_n(t) = \sqrt{n}(\hat{F}_n(t) - t)$ converges weakly to a Gaussian process $Y(t)$ with zero mean and covariance function*

$$\rho(s,t) = \min(s,t) - st - J_1(s)J_1(t) - \frac{1}{2}J_2(s)J_2(t) - \frac{1}{g}J_3(s)J_3(t), \tag{4.2}$$

where $J_1(t) = \phi(\Phi^{-1}(t))$, $J_2(t) = \Phi^{-1}(t)J_1(t)$, $J_3(t) = [(\Phi^{-1}(t))^2 - 1]J_1(t)$, and $s, t \in [0,1]$; the constant g is given by

$$g = 6 + 8c_1 + c_2 - c_1^2 - a(\lim_{n\to\infty} n^{-1}U^tU)^{-1}a^t. \tag{4.3}$$

The existence of the limit in (4.3) is assured by assumption (c) above. The proof of Theorem 4.1 is given in the Appendix.

Comment. Hinkley (1975) obtained the asymptotic variance-covariance matrix of $\sqrt{n}(\hat{\theta} - \theta)$ for the one-sample problem when $\lambda = 0$ and where there is no regression involved, so that g then equals 6. (There is a misprint in his derivation because the asymptotic covariance of $\hat{\mu}$ and $\hat{\nu}$ should be $2\mu(\nu + \mu^2)/3$, instead of $2\mu(\nu + \mu^2)$.) In Theorem 4.1 we have given the asymptotic variance-covariance matrix for the more general case of linear models.

4.2 The case $\lambda \neq 0$

For $\lambda \neq 0$, the integrals involved in the Fisher information matrix are not tractable (see the Appendix), so the asymptotic variance-covariance matrix of $\hat{\theta}$ was examined numerically. It appears that, under mild conditions on the model matrix X, maximum likelihood estimates of the parameters in model (1.3) are again asymptotically normal, for general λ values, and have variance-covariance matrices with the usual Fisher structure. Concerning the empirical process $\hat{Y}_n(t) = \sqrt{n}(\hat{F}_n(t) - t)$, we conjecture that for general λ, σ and β, this can be approximated by a Gaussian process $Y_G(t)$ with zero mean and covariance function

$$\rho_G(s,t) = \min(s,t) - st - \Psi_G^t(s)\Gamma_G\Psi_G(t),\ 0 \leq s,t \leq 1, \tag{4.4}$$

where Γ_G is a $(p+2) \times (p+2)$ matrix and $\Psi_G(t)$ is a $(p+2) \times 1$ column vector function of t. Both Γ_G and $\Psi_G(t)$ depend on n. Expressions for Γ_G and $\Psi_G(t)$ are given in the Appendix.

The next theorem shows that $\rho_G(s,t)$, for many situations occurring in practice, can be well approximated by $\rho(s,t)$.

Theorem 4.2 *Let g_n be the quantity given in (2.14) but using the true parameter values. If (i) $\lambda \to 0$, or (ii) $\sigma \to 0$, or (iii) $\beta_1 \to +\infty$, then the covariance function $\rho_G(s,t)$ of (4.4) converges (pointwise) to the covariance function $\rho(s,t)$ of (4.2) with g in (4.2) replaced by g_n. In general, this pointwise convergence holds if $|(\beta_1 + 1/\lambda)/\sigma| \to \infty$. If $n \to \infty$ is added as a condition, then this pointwise convergence holds without modification.*

The proof of Theorem 4.2 is given in the Appendix.

5 Accuracy of the tests

In this section we discuss the accuracy of the tests given in Section 2. These tests use the points given in Table 1, which are asymptotic points for distributions corresponding to $\lambda = 0$. There are two issues to consider: (a) the accuracy of these points when $\lambda \neq 0$, and (b) the accuracy of the tests in practice, when the asymptotic points are used with finite samples.

Theorem 4.2 suggests that, even when $\lambda \neq 0$, it will often be the case that $\rho_G(s,t)$ can be approximated closely by $\rho(s,t)$, and then Table 1 points might be accurate for practical purposes. This was first studied by comparing $\rho_G(s,t)$ to $\rho(s,t)$ numerically. The results showed that $\rho_G(s,t) \approx \rho(s,t)$ when L of (2.1) can be well approximated by l_{BC} of (2.5); as was discussed in Section 2, this is frequently assumed to be the case, and indeed occurs commonly in practice.

More importantly, to assess the accuracy of the tests, the exact significance levels corresponding to points in Table 1 were calculated from the correct asymptotic distributions using $\rho_G(s,t)$, for a range of parameter values λ, μ and σ. A small sample of results is given in Table 5. These are for statistic W^2, and for the model with no regression, so that $g = 6$ and $\beta_1 = \mu$. The upper percentiles of the asymptotic distribution of W^2, taken from Table 1, and their significance levels, are given in column 1. Recall that these are calculated using $\rho(s,t)$. The next four columns give the values of λ, μ and σ, and the significance levels when the points in column 1 are inserted into the correct asymptotic distribution using $\rho_G(s,t)$. For these examples, $\delta = (\mu + 1/\lambda)/\sigma$; recall that, if $\lambda > 0$ and $\delta = \infty$, or if $\lambda < 0$ and $\delta = -\infty$, Table 1 would be exactly correct. It is clear from Table 5 that Table 1 can be used to give excellent approximations to asymptotic percentage points of W^2, even when δ is far from its limit ∞ or $-\infty$. Similar results hold for A^2.

The next question which arises is how well the asymptotic points approximate the correct points for finite sample size n. In many goodness-of-fit situations it has been verified (see, for example, Stephens, 1986) that the points for finite n, for W^2 and A^2 and other members of the Cramér-von Mises family, converge rapidly to the asymptotic points, and the situation

Table 5: A comparison of asymptotic significance levels for various λ, μ and σ values based on statistic W^2.

Upper Percentiles (α levels)	Parameter Values			
	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = -0.5$
	$\mu = 10$	$\mu = 0.0$	$\mu = .5$	$\mu = -13$
	$\sigma = 0.5$	$\sigma = 0.8$	$\sigma = 1.0$	$\sigma = 1.0$
	$(\delta = 24)$	$(\delta = 3.125)$	$(\delta = 2.167)$	$(-\delta = 15)$
0.0428 (0.50)	0.500	0.528	0.521	0.500
0.0663 (0.20)	0.201	0.198	0.192	0.201
0.0836 (0.10)	0.100	0.098	0.088	0.100
0.1007 (0.05)	0.051	0.048	0.040	0.050
0.1406 (0.01)	0.010	0.009	0.006	0.010

appears to be the same for the present problem. Linnet (1988) empirically studied the use of the Anderson-Darling statistic A^2 and the Cramér-von Mises statistic W^2 to test for normality of the power-transformed data in one-sample problems. Linnet concluded that the null distributions of A^2 and W^2 depend neither on the transformation parameter λ nor on the mean μ and variance ν. A table was provided for A^2 and W^2 for finite samples in which the asymptotic critical points were obtained by extrapolation.

The accuracy of the asymptotic points in Table 1 was investigated here by another method, using the tree data of Example 2 and the biological data of Example 3 in the following simulation study. Consider the tree data. The value of W^2 is 0.0450 and using the asymptotic points the P-value is 0.590. The accuracy of this P-value was examined by taking the estimates of the parameters as the true values and simulating new samples based on this model. The Box-Cox transformation procedure was then applied to each simulated sample and the EDF statistics calculated. The fraction of W^2 values which exceeded 0.0450 gives the empirically derived P-value. This was repeated for the statistic A^2 and the whole experiment repeated again for the biological data. Table 6 gives a comparison between the P-values of the data and the empirical P-values. It can be seen that they are very close, and give evidence that the asymptotic points in Table 1 can be used safely for samples of reasonable size (we suggest $n > 20$).

Table 6

Example		Data $P-$value	Empirical $P-$value
2	$P(W^2 > 0.0450)$	0.5895	0.5833
2	$P(A^2 > 0.2925)$	0.6282	0.6100
3	$P(W^2 > 0.0281)$	0.8700	0.8850
3	$P(A^2 > 0.1974)$	0.8848	0.9117

6 Summary

In summary, tests have been given to assess the normality of a linear model fitted to values obtained by a Box-Cox transformation. The effect of estimating several parameters is contained in one critical parameter g, which must be used to enter Table 1. The points given in this Table are correct asymptotically for $\lambda = 0$, but will also give very good approximations to correct points in most circumstances when $\lambda \neq 0$. They are also good approximations to the correct points for finite samples of reasonable size.

This work was supported by the Natural Science and Engineering Research Council of Canada, and the authors express their thanks for this support.

Appendix

A Proof of Theorem 4.1

Proof of (1). Let $\mu = (\mu_1, \ldots, \mu_n)^t = X\beta$. When $\lambda = 0$, $W_i = Y_i(0) = \log Y_i \sim N(\mu_i, \nu)$. Denote $dY_i(\lambda)/d\lambda$ by $\dot{Y}_i(\lambda)$ and $d^2Y_i(\lambda)/d\lambda^2$ by $\ddot{Y}_i(\lambda)$, then $\dot{Y}_i(0) = W_i^2/2$, $\ddot{Y}_i(0) = W_i^3/3$. Straightforward calculations show that the inverse of the Fisher information matrix for $\theta = (0, \beta^t, \nu)^t$ is given by

$$\Gamma_n = \begin{pmatrix} A_n & B_n \\ B_n^t & C_n \end{pmatrix}^{-1},$$

where

$$n^{-1}A_n = (4\nu)^{-1}(7\nu^2 + 10\nu\beta^t(n^{-1}X^tX)\beta + n^{-1}1_n^t\mu^4),$$

$$n^{-1}B_n = (-D/2 - (n^{-1}X^t\mu^2)^t/(2\nu),\ -\beta_1/\nu),$$

$$n^{-1}C_n = \begin{pmatrix} (n^{-1}X^tX)/\nu & 0 \\ 0^t & 1/(2\nu^2) \end{pmatrix},$$

where $D^t = (1, 0, \ldots, 0)^t$ is a $p \times 1$ vector with its first component equal to 1 and all the other components equal to 0. Therefore, as $n \to \infty$, $n^{-1}\Gamma_n \to \Gamma$ as desired.

Proof of (2). The proof is based on Loynes (1980). In the present case, the null hypothesis $H_n(\gamma)$ in Loynes (1980) specifies nothing and all the parameters $\theta = (\lambda, \beta^t, \nu)^t$ are to be estimated. Since $Y_i(0) = \log Y_i \sim N(\mu_i, \nu)$, all of the expectations needed to form the Fisher information matrix can be found exactly. However, to make the proof more readable, it is assumed that $\beta = 0$ and $\nu = 1$. Then the inverse of the asymptotic variance-covariance matrix for $\theta = (0, 0^t, 1)^t$ is found to be

$$\Gamma = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{7}{6} & 0 & 0 \\ 0 & 0^t & G^{-1} & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

where $G^{-1} = (\lim_{n\to\infty} n^{-1}U^tU)^{-1}$, $X = (1_n\ U)$. It is readily checked that assumptions $A1$ and $A2$ of Loynes (1980) are satisfied naturally by model (1.3). Assuming assumption (D), it can be checked that assumptions $A4$ and $A5$ of Loynes are also satisfied; Assumptions $A7$ and $A9(b)$ of Loynes can be checked by direct calculations. Therefore, by Theorem 1 of Loynes (1980), the estimated empirical process $\hat{Y}_n(t)$ of (4.1) converges weakly to a Gaussian process $Y(t)$. By Corollary 1 of Loynes (1980), the mean of $Y(t)$ is zero and the covariance function of $Y(t)$ is

$$\rho(s,t) = \min(s,t) - st - \Psi^t(s)\Gamma\Psi(t),$$

where $\Psi(t)$ is found to be

$$\Psi(t) = (-\frac{1}{2}[\Phi^{-1}(t)]^2 J_1(t),\ J_1(t)D^t,\ \frac{1}{2}J_2(t))^t, \qquad (s,t \in [0,1]).$$

Direct computations then show that the expression given in (4.2) follows for the case where $g = 6$. □

B Expressions for Γ_G and $\Psi_G(t)$

The $(p+2)\times(p+2)$ matrix Γ_G and the $(p+2)\times 1$ function $\Psi_G(t)$ studied in Theorem 4.2 are given below.

Let I_n be the Fisher information matrix for a random sample $Y_1,\ldots,Y_n$ from model (1.3). Then

$$\Gamma_G = \left(\frac{1}{n}I_n\right)^{-1},$$

and for $\lambda > 0$, I_n has the following components:

$$\begin{aligned}
E\left\{-\frac{\partial^2 L}{\partial\lambda\partial\lambda}\right\} &= \sum_{i=1}^{n}\left\{\frac{1}{\nu}J_{1i} - \frac{\phi(\delta_i)\Phi(\delta_i)(\delta_i - 2\lambda\sqrt{\nu}) + \phi^2(\delta_i)}{\nu\lambda^4\Phi^2(\delta_i)}\right\}, \\
E\left\{-\frac{\partial^2 L}{\partial\lambda\partial\beta}\right\} &= -\frac{1}{\nu}X^tE\{\dot{Y}(\lambda)\} + X^t diag\left(\frac{\delta_i\phi(\delta_i)\Phi(\delta_i) + \phi^2(\delta_i)}{\nu\lambda^2\Phi^2(\delta_i)}\right)1_n, \\
E\left\{-\frac{\partial^2 L}{\partial\lambda\partial\nu}\right\} &= -\frac{1}{\nu^2}E\{(Y(\lambda) - X\beta)^t\dot{Y}(\lambda)\} - \sum_{i=1}^{n}\frac{\phi(\delta_i)\Phi(\delta_i)(\delta_i^2 - 1) + \delta_i\phi^2(\delta_i)}{2\nu\sqrt{\nu}\lambda^2\Phi^2(\delta_i)}, \quad \text{(B.1)}
\end{aligned}$$

$$
\begin{aligned}
E\left\{-\frac{\partial^2 L}{\partial\beta\partial\beta^t}\right\} &= \frac{1}{\nu}X^tX - \frac{1}{\nu}X^t diag\left(\frac{\delta_i\phi(\delta_i)\Phi(\delta_i) + \phi^2(\delta_i)}{\nu\Phi^2(\delta_i)}\right)X, \\
E\left\{-\frac{\partial^2 L}{\partial\beta\partial\nu}\right\} &= X^t diag\left(\frac{\phi(\delta_i)}{\nu\sqrt{\nu}\Phi(\delta_i)} + \frac{\phi(\delta_i)\Phi(\delta_i)(\delta_i^2-1) + \delta_i\phi^2(\delta_i)}{2\nu\sqrt{\nu}\Phi^2(\delta_i)}\right)1_n, \\
E\left\{-\frac{\partial^2 L}{\partial\nu\partial\nu}\right\} &= \frac{n}{2\nu^2} - \sum_{i=1}^{n}\left\{\frac{\delta_i\phi(\delta_i)}{\nu^2\Phi(\delta_i)} - \frac{\delta_i\phi(\delta_i)\Phi(\delta_i)(\delta_i^2-3) + \delta_i^2\phi^2(\delta_i)}{4\nu^2\Phi^2(\delta_i)}\right\},
\end{aligned}
$$

where $diag(d_i)$ denotes $n \times n$ diagonal matrices with d_i as the $(i,i)^{th}$ element, 1_n denotes an $n \times 1$ column vector of 1's. The expressions for the case where $\lambda < 0$ can be obtained by replacing δ_i by $-\delta_i$, except for $E\{-\partial^2 L/\partial\beta\partial\nu\}$, where the whole expression also needs to be multiplied by -1. The J_{1i}'s are given by, for $\lambda > 0$,

$$
\begin{aligned}
J_{1i} &= E\{\dot{Y}_i^2(\lambda) + (Y_i(\lambda) - \mu_i)\ddot{Y}_i(\lambda)\} \\
&= (\lambda^4\Phi(\delta_i))^{-1}\int_{-\delta_i}^{+\infty}\phi(v)[\{(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v\}^2 \\
&\quad +\lambda\sigma v(1+\lambda\mu_i+\lambda\sigma v)\log^2(1+\lambda\mu_i+\lambda\sigma v) \\
&\quad -2\lambda\sigma v\{(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v\}]\,dv, \qquad \text{(B.2)}
\end{aligned}
$$

where $\mu_i = x_i^t\beta$, $\delta_i = (\mu_i + 1/\lambda)/\sqrt{\nu}$, $\dot{Y}_i(\lambda)$ and $\ddot{Y}_i(\lambda)$ are the first and second derivatives of $Y_i(\lambda)$ with respect to λ, respectively; for the case $\lambda < 0$, the above integrals should be done for the range $-\infty$ to $-\delta_i$ and $\Phi(\delta_i)$ should be replaced by $\Phi(-\delta_i)$.

Similarly, $E\{\dot{Y}(\lambda)\}$ has components J_{2i} given by, for $\lambda > 0$,

$$
\begin{aligned}
&J_{2i} = E\{\dot{Y}_i(\lambda)\} \\
&= (\lambda^2\Phi(\delta_i))^{-1}\int_{-\delta_i}^{+\infty}\phi(v)[(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v]\,dv, \quad \text{(B.3)}
\end{aligned}
$$

and $E\{(Y(\lambda) - X\beta)^t\dot{Y}(\lambda)\}$ has components J_{3i} given by, for $\lambda > 0$,

$$
\begin{aligned}
&J_{3i} = E\{(Y_i(\lambda) - \mu_i)\dot{Y}_i(\lambda)\} \\
&= (\lambda^2\Phi(\delta_i))^{-1}\int_{-\delta_i}^{+\infty}\sigma\phi(v)v[(1+\lambda\mu_i+\lambda\sigma v)\log(1+\lambda\mu_i+\lambda\sigma v) - \lambda\mu_i - \lambda\sigma v]\,dv. \quad \text{(B.4)}
\end{aligned}
$$

In the case where $\lambda < 0$, the above two integrals should be done for the range $-\infty$ to $-\delta_i$ and $\Phi(\delta_i)$ should be replaced by $\Phi(-\delta_i)$.

For function $\Psi_G(t)$, there is

$$\Psi_G(t) = \frac{1}{n}\sum_{i=1}^{n} \Psi^{(ni)}(t),$$

where $\Psi^{(ni)}(t)$ is a $(p+2)\times 1$ column vector function with the following components, where $j = 1, \ldots, p$ corresponds to the components associated with β:

$$\begin{aligned}
\Psi_1^{(ni)}(t) &= \begin{cases} -(\sigma\lambda^2\Phi^2(\delta_i))^{-1}[\phi(w_i)\Phi(\delta_i)\{(1+\lambda\mu_i+\lambda\sigma w_i)\log(1+\lambda\mu_i & \\ +\lambda\sigma w_i) - \lambda\mu_i - \lambda\sigma w_i\} + \phi(\delta_i)\Phi(w_i) - \phi(\delta_i)], & \text{if } \lambda > 0, \\ -(\sigma\lambda^2\Phi^2(-\delta_i))^{-1}[\phi(v_i)\Phi(-\delta_i)\{(1+\lambda\mu_i+\lambda\sigma v_i)\log(1+\lambda\mu_i & \\ +\lambda\sigma v_i) - \lambda\mu_i - \lambda\sigma v_i\} - \phi(-\delta_i)\Phi(v_i)], & \text{if } \lambda < 0, \end{cases} \\
\Psi_{2j}^{(ni)}(t) &= \begin{cases} \{x_{ij}/(\sigma\Phi^2(\delta_i))\}\{\phi(w_i)\Phi(\delta_i) + \phi(\delta_i)\Phi(w_i) - \phi(\delta_i)\}, & \text{if } \lambda > 0, \\ \{x_{ij}/(\sigma\Phi^2(-\delta_i))\}\{\phi(v_i)\Phi(-\delta_i) - \phi(-\delta_i)\Phi(v_i)\}, & \text{if } \lambda < 0, \end{cases} \qquad \text{(B.5)} \\
\Psi_3^{(ni)}(t) &= \begin{cases} (2\sigma^2\Phi^2(\delta_i))^{-1}\{\phi(w_i)w_i\Phi(\delta_i) - \delta_i\phi(\delta_i)\Phi(w_i) + \delta_i\phi(\delta_i)\}, & \text{if } \lambda > 0, \\ (2\sigma^2\Phi^2(-\delta_i))^{-1}\{\phi(v_i)v_i\Phi(-\delta_i) + \delta_i\phi(-\delta_i)\Phi(v_i)\}, & \text{if } \lambda < 0, \end{cases}
\end{aligned}$$

where $w_i = w_i(t) = \Phi^{-1}(1+\Phi(\delta_i)(t-1))$, $v_i = v_i(t) = \Phi^{-1}(t\Phi(-\delta_i))$, and $t \in [0,1]$.

C Proof of Theorem 4.2

Define $\delta = (\beta_1 + 1/\lambda)/\sigma$ and $\eta_i = u_i^t(\beta_2, \ldots, \beta_p)^t/\sigma$, where u_i^t is the i^{th} row of U. Then as $\delta \to \infty$, $\phi(\delta_i) \to 0$ and $\Phi(\delta_i) \to 1$. Now let $\delta \to \infty$, we have an expansion for the function $\Psi_G(t)$ given by

$$\Psi_G(t) = -\{J_1(t)/\sigma\}(e_n(t), D, \Phi^{-1}(t)/(2\sigma))^t, \qquad \text{(C.6)}$$

where $D = (1, 0, \ldots, 0)$ is a $1 \times p$ vector and $e_n(t)$ is given by

$$\begin{aligned} e_n(t) &= \frac{1}{\lambda^2} + \frac{\sigma}{\lambda}\{\delta\log(\lambda\sigma\delta) - \delta + \log(\lambda\sigma\delta)\Phi^{-1}(t) \\ &\quad + \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{\delta^k k(k+1)}\frac{1}{n}\sum_{i=1}^{n}(\eta_i + \Phi^{-1}(t))^{k+1}\}. \qquad \text{(C.7)} \end{aligned}$$

Similarly, let $\delta \to \infty$, we can have an expansion for Γ_G given by

$$\sigma^2\Gamma_G = \begin{pmatrix} a_n & b_n & c_n \\ b_n & (X^tX)/n & 0 \\ c_n & 0^t & 1/(2\sigma^2) \end{pmatrix}^{-1}, \tag{C.8}$$

where b_n is the result of replacing $\Phi^{-1}(t)$ in (C.7) by ε_i and taking expectations; c_n is the result of multiplying (C.7) by $\Phi^{-1}(t)$, then replacing $\Phi^{-1}(t)$ by ε_i and taking expectations; and a_n can be obtained in a similar but more involved fashion. The key thing is that with the above expansions, $\Psi^t_G(s)\Gamma_G\Psi_G(t)$ turns out to depend on δ and the η_i's, and not to depend on λ, σ and β_1. Using elementary row and column reductions to simplify $\Psi^t_G(s)\Gamma_G\Psi_G(t)$ into an expression in terms of $e_n(s)$, $e_n(t)$, a_n, b_n and c_n, and substituting the above expansions into this expression will lead to the desired result. Details of the algebra are available from the authors. □

D Calculation of Percentage Points

The asymptotic distribution of W^2 is determined by the eigenvalues of $\rho(s,t)$ and the asymptotic distribution of A^2 is determined by the eigenvalues of $\rho(s,t)/\{s(1-s)t(1-t)\}^{1/2}$. In general, let the covariance function of a integral type statistic T be $k(s,t)$. Then the limiting distribution of T has the form $\sum_{i=1}^{\infty} \lambda_i\chi^2_i$, where the χ^2_i are independent chi-square random variables on 1 degree of freedom, and the λ_i's are the eigenvalues of the integral equation

$$\lambda f(s) = \int_0^1 k(s,t)f(t)dt.$$

For the general theory behind the above statements, see Durbin (1973). In this paper, the above equation is discretized into

$$\lambda f_i = \frac{1}{m}\sum_{j=1}^{m} k\left(\frac{i-0.5}{m}, \frac{j-0.5}{m}\right) f_j, \ (i = 1, \ldots, m)$$

for a large integer m ($m = 100$ is used in this paper) and can be solved for eigenvalues $\hat{\lambda}_i$ ($i = 1, \ldots, m$). Then the distribution of $\sum_{i=1}^{\infty} \lambda_i\chi^2_i$ can be approximated by $\sum_{i=1}^{m} \hat{\lambda}_i\chi^2_i + \tau\chi^2_{m+1}$, where χ^2_{m+1} is a chi-square random variable on 1 degree of freedom and independent of the χ^2_i $(i = 1, \ldots, m)$, and τ is found by making

$$\int_0^1 k(t,t)dt = \sum_{i=1}^{\infty} \lambda_i = (\sum_{i=1}^{m} \hat{\lambda}_i) + \tau$$

true. For example, for W^2, $\sum_{i=1}^{\infty} \lambda_i = 0.0492385$, $\sum_{i=1}^{m} \hat{\lambda}_i = 0.0492413$ so $\tau = -2.8\text{e-}06$. Finally, the percentage points are found using the numerical Fourier inversion method of Imhof (1961).

References

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations (with discussion). *Journal of Royal Statistical Society*, **B, 26**, 211-252.

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*, Regional Conference Series in Applied Mathematics, 9. Philadelphia: SIAM.

Hinkley, D.V. (1975). On Power Transformations to Symmetry. *Biometrika*, **62**, 101-112.

Imhof, J.P. (1961). Computing the Distribution of Quadratic Forms in Normal Variables. *Biometrika*, **48**, 419-426.

Linnet, K. (1988). Testing Normality of Transformed Data. *Applied Statistics*, **37**,180-186.

Loynes. R. M. (1980). The Empirical Distribution Function of Residuals from Generalised Regression. *Annals of Statistics*, **8**, 285-298.

Poirier, D.J. (1978). The Use of the Box-Cox Transformation in Limited Dependent Variable Models. *Journal of American Statistical Association*, **73**, 284-287.

Ryan, T., Joiner, B. and Ryan, B. (1976). *Minitab Student Handbook*. Duxbury Press, North Scituate, Massachusetts.

Stephens, M.A. (1986). Chapter 4 in *Goodness-of-Fit Techniques*, (R. B. D'Agostino, and M. A. Stephens, eds). New York: Marcel Dekker.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
EDF Tests for Normality in Linear Models After A Box-Cox Transformation

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER
472

7. AUTHOR(s)
G. Chen, R. Lockhart & Michael A. Stephens

8. CONTRACT OR GRANT NUMBER(s)
N0025-92-J-1264

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Stanford University
Stanford, CA 94305-4065

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR-042-267

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Statistics & Probability Program
Code 111

12. REPORT DATE

13. NUMBER OF PAGES
34

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Keywords: LINEAR REGRESSION, NON-LINEAR REGRESSION, MAXIMUM LIKELIHOOD ESTIMATION, TRANSFORMATIONS TO NORMALITY

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See Reverse Side

EDF Tests for Normality in Linear Models after a Box-Cox Transformation

Gemai Chen, Richard Lockhart and Michael A. Stephens
Simon Fraser University, B.C. Canada

Summary

The Box-Cox transformation procedure has been used extensively in data analysis, for example in regression, where the response variable is subjected to a suitable power transformation so that the standard normal-theory linear regression models can be fitted to the transformed values. In this paper, distribution theory is developed for a family of EDF statistics, including the Anderson-Darling statistic A^2 and the Cramér-von Mises statistic W^2, so that these statistics can be used to test for normality in the linear model after applying the Box-Cox transformation. A table of asymptotic critical points is given for A^2 and W^2, and numerical examples are given to illustrate the use of the table.